Evaluation of Performance of VDSR Super Resolution on Real and Synthetic Images

D. Vint, G. Di Caterina, J. J. Soraghan
Centre for Signal & Image Processing
University of Strathclyde
Glasgow, United Kingdom
{david.vint, gaetano.di-caterina, j.soraghan}@strath.ac.uk

R. A. Lamb, D. Humphreys
Airborne & Space Systems Division
Leonardo MW Ltd
Edinburgh, United Kingdom
{david.humphreys, robert.lamb}@leonardocompany.com

***Abstract*—This paper presents an evaluation of the suitability of the VDSR Single Image Super Resolution architecture, to increase the spatial resolution of lower quality images. For this aim, two sets of tests are performed. The former being on real life images to determine the networks ability to improve low resolution images. The second test is performed on images of a resolution chart, and therefore synthetic. This is to analyse the frequency response of the network. For each test, three metrics are used to assess image quality. These are the PSNR, SSIM and MTF values. Experimental results show that the VDSR network is able to increase the quality of the images within the first test in all three metrics, therefore showing that the network is suitable for super resolution. The second test provides more information on the limitations of the network when given a high contrast image, and the resulting ringing effects it can create. Therefore results in PSNR/SSIM values are not improved over the low resolution images, however they have a higher MTF curve as well as more visually sharp images.**

***Index Terms*—Diffraction, Deep Learning, VDSR, SSIM, MTF**

I. Introduction

The effect of diffraction on imaging systems can result in poor quality images with limited spatial resolution. This is especially true in situations where a long distance image is being captured, as for example in air borne imaging systems. Where the use of a long focal length can scale the effects of diffraction when combined with a finite aperture size.

It is therefore required that these affected images undergo post processing, as to increase their spatial resolution. Super Resolution is a process that aims to achieve this. With the aim of constructing high spatial resolution images from single or multiple low-resolution ones, where Single Image Super Resolution (SISR) is especially prevalent. This task of producing a high resolution image is ill-posed, as given any low-resolution image, a great number of possible high-resolution versions of it could exist. This task of SISR has recently been mostly achieved with deep learning solutions, as they are able to provide a high quality result when compared to traditional image processing techniques.

The Very Deep Super Resolution (VDSR) architecture introduced by Kim et al. [1] was the first deep learning SISR technique that extended their methodology to adopt the idea that deeper networks give greater performance. This new architecture made use of the fact that when performing super resolution, the direct pixel to pixel mapping that had been used in the past was overly complex and resulted in an extremely slow convergence rate. To overcome this problem, the VDSR architecture learns the residual between low-resolution and high-resolution images as opposed to the direct mapping. This allows the network to be far deeper, whilst still maintaining a suitable convergence rate.

In this paper, we investigate the performance of the VDSR network. The assessment is carried out with two separate approaches. Firstly, we asses the network against a subset of images extracted from the dataset used to train the network, as well as some external aerial images obtained from a separate dataset. Secondly, the network is assessed against a set of images of a resolution chart, hence synthetic, where the images resolution is affected by the diffraction, determined by the camera system.

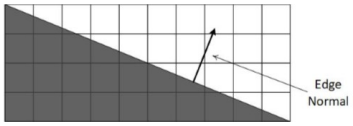
For both approaches, the resulting 'super-resolved' images are evaluated with three separate metrics. These are the Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) and the Modulation Transfer Function (MTF), where the latter is a non-traditional image improvement metric. The MTF is traditionally used to evaluate the spatial resolution of camera systems, and is adopted in our experiments to evaluate how the VDSR network affects the frequencies of the given images. The purpose of this evaluation is to determine what it is learned by the VSDR deep convolutional network during the training phase, especially in the context of images that were not present during training and are specially chosen to evaluate the network's performance.

For a fair evaluation of the network, the provided architecture was re-trained on a known dataset. This way, the results could be easily evaluated. For this purpose, the iaprtc12 [2] image dataset was used, as it is a publicly available dataset comprising of 20,000 still natural images. As well as images from such a dataset, some additional images were also obtained from other sources. These images were aerial shots taken from the DOTA [3] aerial image dataset, and used to asses the networks ability to enhance an aerial image that may then be used for object identification and recognition.

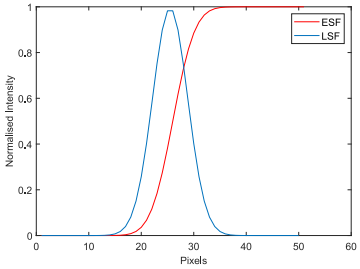
To obtain the synthetic images, a Nikon D3100 camera was used to capture still images of an ISO 12233 resolution chart [4] that had been printed off and attached to a vertical surface.

D.V. acknowledged support from the UK EPSRC and from Leonardo MW Ltd, Edinburgh.

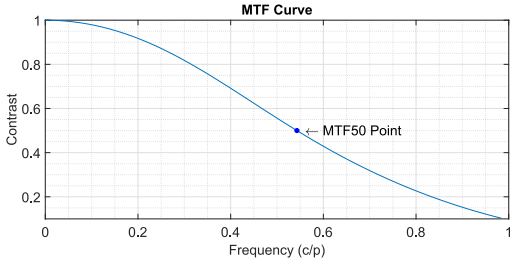
(a) An edge transition from black to white at an angle.



(b) Slanted edge falling on different pixel areas of the camera sensor, resulting in small differences between individual pixels.



(c) Methodology to transition from Edge Spread Function (Red) to Line Spread Function (Blue) via differentiation. Then finally to the MTF curve via the Fourier Transform



(d) Methodology to transition from Edge Spread Function (Red) to Line Spread Function (Blue) via differentiation. Then finally to the MTF curve via the Fourier Transform

Fig. 1: Separate Elements of MTF acquisition process.

From this resolution chart, the MTF can be calculated in a number of different ways, and it could then be used to analyse the frequency performance of the VDSR network.

This paper is organised as follows; Section II discusses related works in SISR. Section III describes the training approach. Section IV discusses the three quality assessment metrics used for analysis. Section V describes the experimental procedure carried out well as the results. Finally, Section VI presents a conclusion, with some final comments on the results, and their validity.

II. Related Works

Approaches to super resolution via traditional methods can be separated into four categories [5]. These are prediction models, edge based methods, image statistical methods and patch based models. An in depth review of these models is presented in [6].

The introduction of convolutional neural networks (CNN) has brought considerable enhancement to the field of super resolution. The first CNN proposed for such a task was 'SRCNN' from Dong et al. [5], which successfully applied a convolutional network to the problem of super resolution. After this first attempt, a great number of new SISR architectures for super resolution were proposed, including the SRResNet architecture [7], which adapted the ResNet architecture [8] for super resolution. This was then further improved upon by Lim et al. [9], who realised that the removal of certain elements from the ResNet block was more suited the task of super resolution.

As well as these networks, a great number of other methods have been developed, which has prompted the creation of an annual super resolution Deep Learning-based competition, NTIRE [10]. The purpose of this competition is to determine a performance benchmark for the state of the art SISR systems.

III. Very Deep Super Resolution Architecture

The VDSR network [1] was chosen as the network to be analysed due to its ease of use in the Matlab environment, where a pre-defined example already existed within the Mathworks documentation [11]. Using this, the code for the VDSR network was recreated locally, allowing the network to be re-trained on a larger dataset than the pre-trained network provided by Mathworks.

The overall architecture of the VDSR Network can be reviewed in the original paper [1]. It comprises 20 weight layers, in which all are identical except for the first and last. The first layer takes in the single channel luminance of an input image. The bulk of the network then consists of 18 convolutional layers, each of which has 64 3×3 filters, and each convolutional layer is followed by a rectified linear unit (ReLu). The final layer then reconstructs the desired residual image with a $3 \times 3 \times 64$ filter. This final residual image can then be combined with the low-resolution input to provide the resulting super-resolved output.

To train the network, 1,477 high-resolution images were extracted from the iaprtc12 dataset. To generate the corresponding low resolution images, the high-resolution images were down sampled by a scale factor of 2, 3 or 4. Once downscaled, they were then up-scaled to their original resolution, resulting in a suitably distorted low resolution image dataset. From each of these images, 64 41×41 image patches were extracted, some of which were then randomly rotated by $90°$. This patch extraction and rotation allows the network to train faster as well as ensuring a suitable diversity in its training data. The training parameters of the network were kept the same as those provided in the Matlab example.

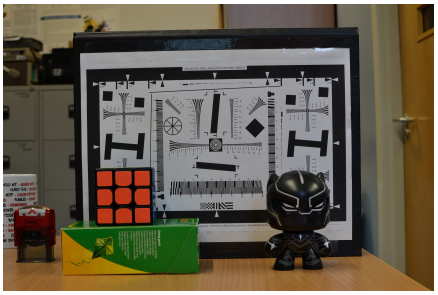
IV. Quality Assessment

A. Peak Signal to Noise Ratio

The Peak Signal to Noise Ratio, (PSNR) is the most common image comparison metric, and evaluates the noise power

(a) Example image from the iaprtc12 dataset with additional slanted edge on lower right corner.

(b) Example image from the DOTA datset containing aerial images. With additional slanted edge on lower right corner.



(c) Lab-based image of ISO 12233 Resolution chart, taken with a DSLR camera.

Fig. 2: Example images used for testing.

present within a signal. For the case of Super resolution, this noise is manifested by a lack of detail in the low-resolution images. It is calculated alongside the Mean Square Error (MSE), which is a measure of the squared average difference between two images. The MSE is also the loss function used for the VDSR network during training. Therefore, the PSNR is a good metric to determine the quality improvment provided by the network.

B. Structural Similarity Index

The SSIM image metric introduces a different way of comparing two images. Instead of attempting to determine the differences between images, the SSIM makes use of the luminance, contrast and structure of the images to provide the perceived quality difference between two images. Unlike the PSNR, the SSIM metric is based on visible structures within the image. The full description of the SSIM can be found in [12].

C. Modulation Transfer Function

Traditionally, the MTF one of a number of metric used to assess the quality of imaging systems [13] [14]. This can be useful to determine the spatial resolution performance of any given imaging system. To accurately obtain the MTF of a given system, a number of different methods can be utilised. Ideally, the MTF is calculated from the point spread function (PSF), which is a camera's response image to a single point of light. However, the PSF is difficult to obtain in practice. To overcome this, the slanted edge method was developed [15]. This is used to obtain the Edge spread function (ESF). From which can be extracted the Line Spread Function (LSF). Both of which can be seen in Fig. 1c. Where the ESF is obtained through the analysis of a 'slanted edge' (Fig. 1a). This method is effective, as the length of the edge allows the ability to 'super sample' the edge data. As well as this, the slant that is present allows the edge to fall on different areas of the camera's sensor as seen in Fig. 1b. Once the ESF has been extracted, it is differentiated to produce the corresponding LSF. Finally, to obtain the MTF curve, the Fourier Transform of the LSF is taken. This produces a curve that describes the spatial resolution of the image. A key data point that can be extracted from this curve is the MTF50 point. This is the spatial frequency at which the contrast drops to 0.5. This value is the metric that is used for analysis throughout this paper.

An example of a resulting MTF curve can be seen in Fig. 1d, where the y-axis is the contrast, measured on a normalised scale of 0 to 1. The x-axis represents the spatial frequency in Cycles per Pixel.

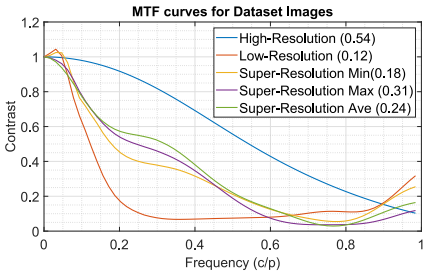
The MTF was calculated with the use of the open source program 'MTF Mapper' [16]. This program has a multitude of different functionalities. For the purposes of this paper, MTF mapper took in a slanted edge as input (Fig. 1a) and produced a corresponding MTF curve as an output (Fig. 1d).

V. Experimental Results

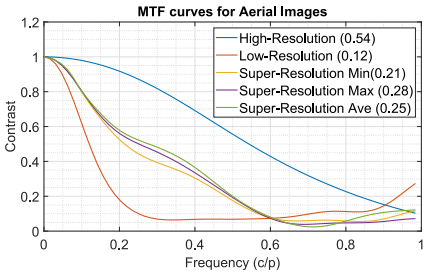
1) Real images: Test one was on a subset of images extracted from the iaprtc12 dataset [2], as well as images taken from the DOTA aerial dataset [3]. Where the images from the iaprtc12 dataset consisted of 616 images, each of which had been pre-processed before being passed through the network to allow the MTF to be calculated for the images before and after enhancement. To achieve this, a slanted edge was overlaid onto the lower right corner of each image (Fig. 2a). As well as the images taken from iaprtc12, 32 aerial images were taken from the DOTA dataset, where these images were also pre-processed so that the slanted edge was present in the lower right hand corner (Fig. 2b).

As well as the addition of the slanted edge to the test images, the low-resolution images to be used as the input to the system were to be generated. This was achieved by downscaling the high-resolution images from the dataset by a factor of 4, which were then subsequently up sampled to regain their original resolution. Once the pre-processing for all of the images was complete, each image was fed into the VDSR network, and the super-resolved images were obtained and PSNR, SSIM and MTF evaluated.

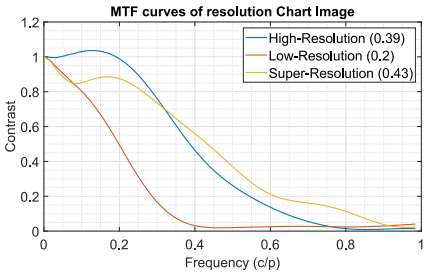
Firstly, the images from the iaprtc12 dataset were reviewed. It was found that for all 616 images, the PSNR and SSIM always improved. This showed that the network was producing suitable residuals that resulted in higher quality super resolved images. To analyse the MTF values of the results, as determined by PSNR and SSIM, the best four images and the worst four images would be analysed in



(a) MTF curves from a number of images taken from the iaprtc12 dataset.



(b) MTF curves from a number of images taken from the DOTA dataset.



(c) MTF curves from a synthetic image of a resolution chart.

Fig. 3: Resulting MTF curves from various image tests. Terms in brackets refer to the corresponding MTF50 values for each curve

greater detail. For each image, the MTF curve was obtained and the MTF50 point was extracted. This resulted in 16 different values. The summary of these results can be seen in TABLE I. As well as the MTF50 values, a plot of the highest, lowest and median MTF curves can be seen in Fig. 3a. Where the MTF of the high-resolution and low-resolution images have also been plotted. The same process was then applied to the 32 aerial images, where, in agreement with the dataset images, the PSNR and SSIM values improved for each. Again, the MTF curve was evaluated for each image, and the MTF50 values were extracted. The results of these plots can be seen in TABLE I. As with the dataset images, the max, min and median MTF50 value curves were plotted alongside the high-resolution and low-resolution MTF curves. This can be seen in Fig. 3b. As can be seen, as well as PSNR/SSIM improvements, the MTF values also increase when the input images are super-resolved with the VDSR network. This shows that the network is not only successfully increasing the overall quality of the image, but also increasing the effective spatial resolution.

TABLE I: MTF curve results.

	'iaprtc12' Images	**'DOTA' Aerial Images**
Minimum MTF50 Value	0.18 c/p	0.21 c/p
Maximum MTF50 Value	0.31 c/p	0.28 c/p
Average MTF50 Value	0.24 c/p	0.26 c/p
Standard Deviation of MTF50 Values	0.031 c/p	0.017 c/p

Low-Res MTF50: 0.12 c/p **High-Res MTF50:** 0.54 c/p

2) Synthetic Images: The second test was on a number of images of a resolution chart, and therefore referred to as synthetic, taken with a Nikon D3100 DSLR camera in a lab environment. The purpose of this test was two fold; firstly, to evaluate how the network responded to images out with the iaprtc12 dataset; secondly, the inclusion of the resolution chart allowed analyses of exactly what the network was attempting to learn, in order to improve the given low resolution image. The images taken were of an ISO 12233 resolution chart. An example of one of these images can be seen in Fig. 2c.

To obtain the low-resolution counterparts of the synthetic images, the F-stop setting of the camera was altered to produce a lower quality image. This reduced the aperture of the lens, therefore increasing the effect of diffraction. This was done instead of down/up sampling to evaluate how well the network performs on real images that have been affected by diffraction. For the final tests, the high and low resolution images were taken with F-stops of F9 and F22 respectively.

The only pre-processing required for the synthetic images was that of cropping. Due to limitations of memory for the network, the full images obtained from the camera (4608×3072 pixels) could not be processed. Therefore, the images were cropped to images of 480×360 pixels, as to match those from within the dataset. For each image, four separate crops were taken, resulting in a total of 16 images. These images were then passed through the VDSR network, and the super-resolved images were obtained. As with the previous test, the PSNR and SSIM were calculated for each of the 16 images. In this case, contradictory of what was found in the first test, the PSNR and SSIM had decreased for every image with exception of one. However, when the MTF of individual images was examined, it was found the the MTF50 had increased above even the high-resolution counterpart. This MTF curve can be seen in Fig. 3c.

The reason for the lack of improvement in PSNR/SSIM is due to the fact that the content of the synthetic images contains a very high contrast. This is vastly different from what the network has seen during training. Therefore, it cannot perform as well. However, this is not to say that it does not make any improvements entirely. This improvement can be seen in the MTF curve, which clearly shows that the spatial frequency response has been improved. As well as the MTF, the image itself can also be analysed to determine how the the network is affecting the high frequency components. A comparison of such a high frequency patch is illustrated in Fig. 4, where it can be seen that the network is able to enhance the pixel intensity of the black, whilst also lowering the pixel intensity of the white, resulting in an image that looks of a higher quality than that of the low-resolution input. However, by inspecting the surrounding pixels of the black lines, it can be seen that a certain amount of ringing occurs. An explanation for this could be due to the Gibbs

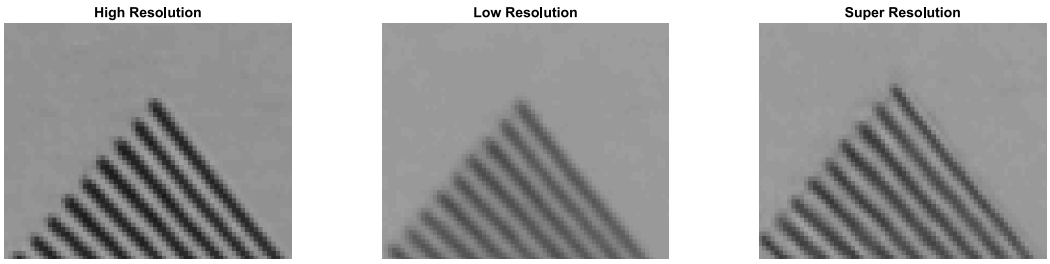


Fig. 4: High frequency patch of resolution chart used to analyse frequency response of VDSR network

phenomenon, where by the neural network is attempting to recreate the perfect periodic pattern, found in the high-resolution image, from the somewhat sinusoidal waves found in the low resolution image. Therefore, due to possible Fourier approximation introduced by the VDSR network, the periodic pattern cannot be perfectly recreated, therefore resulting in the ringing effects seen. It can therefore be appreciated that the resulting super resolved images have lower PSNR/SSIM due to these differences caused by ringing. However, when analysed visually, the super-resolved image does indeed seem to be of a higher contrast than the low resolution input. Therefore, it could be said that in certain cases, the PSNR/SSIM metrics may not be the most suitable, especially if the core concern of the super resolution system is with regard to spatial resolution of high contrast images.

VI. Conclusion

This paper has presented an evaluation of the performance of the VDSR neural network architecture for super resolution. This has been achieved with three different image quality metrics, the PSNR, SSIM and MTF values. Two separate tests were carried out to assess the super resolution ability of the network. Firstly, the network was tested on real life images, taken from the dataset that was used in training, and from an external dataset of aerial images. The results from this test showed that the network was able to improve the given images for all three metrics. This proved that the network is able to increase the spatial resolution of given images, which is useful for when images have been distorted by effects such as diffraction.

The second test that was carried out was on a number of images of a resolution chart. The motivation for this test was to understand what the network was performing on the input images, and the resolution chart allowed a thorough analysis of the output images. It was found that, when analysing a high frequency component of the resolution chart (black and white lines), the network was able to increase the visual contrast of the image, however resulting in worse PSNR/SSIM values due to ringing.

References

[1] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," *CVPR*, vol. abs/1511.04587, 2016.

[2] Imageclef, "IAPR TC-12 Benchmark." https://www.imageclef.org/photodata.

[3] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale Dataset for Object Detection in Aerial Images." https://captain-whu.github.io/DOTA/dataset.html.

[4] Imatest, "ISO 12233 - Resolution and Spatial frequency Responses." http://www.imatest.com/solutions/iso-12233/.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 295–307, Feb 2016.

[6] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image Super-Resolution: A Benchmark," in *Computer Vision - ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. 8692, pp. 372–386, Springer International Publishing, Sept. 2014.

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jul 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.03385, 2015.

[9] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," *CoRR*, vol. abs/1707.02921, 2017.

[10] R. Timofte, S. Gu, J. Wu, L. V. Gool, L. Zhang, and M.-H. Yang, "NTIRE 2018 Challenge on Single Image Super-Resolution: Methods and Results," *NTIRE 2018*, 2018.

[11] Mathworks, "Single Image Super-Resolution Using Deep Learning." https://uk.mathworks.com/help/images/single-image-super-resolution-using-deep-learning.html.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, apr 2004.

[13] G. D. Boreman, *Basic Electro-Optics for Electrical Engineers*, ch. 2, pp. 23–29. SPIE, 1998.

[14] G. D. Boreman, *Modulation Transfer function in Optical and Electro-Optical Systems*, ch. 4, pp. 73–76. SPIE, 2001.

[15] Strolls with my Dog, "The Slanted Edge Method." https://www.strollswithmydog.com/the-slanted-edge-method/.

[16] F. van den bergh, "MTF Mapper." https://sourceforge.net/p/mtfmapper/home/Home/.